Analyzing Interpretability of Summarization Model with Eye-gaze Information

Fariz Ikhwantri, Hiroaki Yamada, Takenobu Tokunaga
Tokyo Institute of Technology, Japan
Meguro, Ôokayama 2-12-1, Tokyo, 152-8550, Japan
ikhwantri.f.aa@m.titech.ac.jp, yamada@c.titech.ac.jp, take@c.titech.ac.jp

Abstract

Interpretation methods provide saliency scores indicating the importance of input words for neural summarization models. Prior work has analyzed models by comparing them to human behavior, often using eye-gaze as a proxy for human attention in reading tasks such as classification. This paper presents a framework to analyze the model behavior in summarization by comparing it to human summarization behavior using eye-gaze data. We examine two research questions: RQ1) whether model saliency conforms to human gaze during summarization and RQ2) how model saliency and human gaze affect summarization performance. For RQ1, we measure conformity by calculating the correlation between model saliency and human fixation counts. For RQ2, we conduct ablation experiments removing words/sentences considered important by models or humans. Experiments on two datasets with human eye-gaze during summarization partially confirm that model saliency aligns with human gaze (RQ1). However, ablation experiments show that removing highly-attended words/sentences from the human gaze does not significantly degrade performance compared with the removal by the model saliency (RQ2).

Keywords: language-model, neural network, interpretability, summarization, eye movement

1. Introduction

Interpretation of deep neural network models has recently drawn much attention in the natural language processing (NLP) community (Doshi-Velez and Kim, 2017; Lipton, 2018; Belinkov et al., 2020). To analyze a model behavior, past studies have proposed various interpretation methods that provide saliency of input words (Simonyan et al., 2013; Ribeiro et al., 2016; Sundararajan et al., 2017; Guan et al., 2019). One research stream to understand the model behavior compares it with human behavior when they solve the same task. Particularly, eye-gaze information has been often used as a surrogate of human behavior relying on the eye-mind assumption (Just and Carpenter, 1980), which claims "…the eye remains fixated on a word as long as the word is being processed." So, the gaze duration on a fixated word indicates the time to process the word. In comparison, the eye-gaze information on input tokes can be a counterpart of saliency provided by the interpretation methods (Sood et al., 2020a; Hollenstein and Beinborn, 2021; Ikhwantri et al., 2023).

The eye movement study has a long history of investigating various human cognitive functions, encompassing tasks like text reading, scene perception, and visual search (Rayner, 1998; Richardson et al., 2007; Rayner, 2009). Particularly, the eye movements during reading activity have been subject to comprehensive investigation (Clifton et al., 2007).

In contrast, eye movement studies on the writing process have been less studied. Carl and Kay (2012) studied eye movement during translation and found the behavior between professional translators and students is different. The professionals read the source text and wrote its translation almost in parallel, while the students did these two phases more interleaving way.

In this study, we investigate behaviors of neural summarization models in terms of human eye-gaze information collected during the human summarization activity. Similar to translation, summarization involves reading and generating a text. Still, in addition, it should identify the core ideas of the source text and generate a coherent short text that covers them. The languages of the source text and its summary are the same, unlike translation.

Recent progress on pre-trained Transformer-based summarization models (Liu and Lapata, 2019; Stiennon et al., 2020; Lewis et al., 2020) improved the model performance by a large margin. Xu and Durrett (2021) investigated the inner-working process of a pre-trained transformer-based summarization model by breaking it down into different parts of models, the pre-trained and fine-tuned stages, and the Transformer components. They also addressed the difference in representation for interpreting the classification and autoregressive generation model. However, they did not consider comparing the model and human behavior.

This paper proposes a framework to analyze neural summarization models through comparison with human summarization behavior. We utilize human eye-gaze and keystroke data as a proxy of human behavior. Through the analysis, we an-

LREC-COLING 2024, pages 939–950
20-25 May, 2024.

swer the following two research questions.

RQ1. Does the input word saliency from interpretation methods in summarization models conform with human eye-gaze features during summarization?

RQ2. How do the model saliency and human visual attention from eye movement affect model performance?

To answer RQ1, we analyze whether the machine looks at the same input elements as humans during summarization. We measure their conformity by calculating the correlation between saliency scores from the summarization models and fixation counts from human eye-gaze data of words in the source texts.

To answer RQ2, we take the ablation approach following DeYoung et al. (2020); Xu and Durrett (2021); at a generation of each word of the summary, the words of interest by the models or humans are removed from the source text, and the quality of the final summary is assessed. If the saliency score and eye-gaze feature represent the importance of words in the source text, the word removal degrades the summary quality.

We conduct experiments using two datasets that include eye-gaze data during human summarization.

2. Related Work

2.1. Interpretation of NLP Models

Interpreting neural networks in classification tasks is defined as assigning an importance score to an input element for the model to output the results (Ancona et al., 2018). For example, in text classification, given an input text of X of n tokens $(x_1, x_2, ..., x_n)$, the model predicts an output $y \in \{c_1, c_2, \ldots c_k\}$, where k is the number of classes. We define the saliency score $\phi(x_i, y)$ for each token $x_i \in X$, which indicates the importance of x_i to classify the text X.

Early studies on analyzing NLP models visualize the input salience for model output in terms of several linguistic properties (Li et al., 2016) in manifold space (van der Maaten and Hinton, 2008). Recently, the gradient-based interpretation method has been popular for input saliency calculation, which was initially introduced in the computer vision field. The saliency scores of pixels in the input image were calculated using backpropagation of the gradient (Simonyan et al., 2013). In the NLP models, saliency scores are calculated for input words, typically represented as a vector $\vec{v_i} \in \mathbb{R}$ at the embedding layer. The saliency $\phi(x_i, y)$ of the input word x_i is calculated as the Euclidean norm of the the gradient of the embedding vector

$$\phi(x_i, y) = \|\nabla_{x_i} f_y(X)\|_2, \quad (1)$$

where X is an input consisting of x_i, and f_y denotes a function corresponding to the task, e.g., text classification. We obtain a saliency vector $[\phi(x_1, y), \phi(x_2, y), ..., \phi(x_n, y)]$, where each element corresponds to the input word.

Feng et al. (2018) assessed neural models' sensitivity to input alterations using gradient-based methods (Sundararajan et al., 2017) in reading comprehension tasks. They reported that the model frequently gained high saliency scores for less important parts of the input texts.

Analyzing the attention layer is also one of the common and convenient methods to interpret deep learning models adopting the attention mechanisum (Bahdanau et al., 2014). The effectiveness of the attention layer for the interpretation method is still controversial (Jain and Wallace, 2019; Wiegreffe and Pinter, 2019; Serrano and Smith, 2019; Vig, 2019; Vashishth et al., 2019; Wiegreffe and Pinter, 2019). DeYoung et al. (2020) and Atanasova et al. (2020) developed benchmarks to evaluate interpretation methods with human annotation. Both studies found that the gradient-based method performs better than the attention-based method. However, other studies still claimed the effectiveness of the attention layer for an explanation of the model performance (Wiegreffe and Pinter, 2019) and the model behavior similar to human eye movement, especially for Transformer-based models (Eberle et al., 2022; Ikhwantri et al., 2023).

Interpreting models for Reading Task An initial study by Hollenstein et al. (2019) compiled a collection of different modalities of cognitive data such as eye-tracking, EEG, and fMRI to evaluate word embedding semantic information. Sood et al. (2020a) focused on the attention layer in deep learning models and eye movements in question-answering (QA) tasks. Hollenstein and Beinborn (2021) compared a language model and human behavior by calculating a correlation between word importance from the model and that from human eye movements. Ikhwantri et al. (2023) conducted a comprehensive investigation of the models for various NLP tasks, such as sentiment analysis, QA, and relation classification. All these studies target NLP tasks involving reading texts.

Interpreting models for Writing Task The autoregressive generation has been a popular technique for neural network-based language generation, where a word is generated according to the probability distribution over the vocabulary at each time step (Alvarez-Melis and Jaakkola, 2017; Vafa

et al., 2021). The probability distribution is determined based on the network state at the previous time step. Unlike most reading tasks that result in a single output, we have a sequence of outputs (words) in generation tasks. Therefore, we have a saliency distribution of input words at each time step, resulting in a saliency distribution matrix instead of a saliency vector.

The interpretation study for sequential generation has been active in the translation task (Alvarez-Melis and Jaakkola, 2017; Vafa et al., 2021; Voita et al., 2021). The apparent application of the saliency score matrix is to analyze the alignment between source and target tokens in Machine Translation (Ding et al., 2019; He et al., 2019). This naturally can be used to calibrate attention models to improve performance (Lu et al., 2022). Recently, hallucination has been analyzed with the model interpretation method (Tang et al., 2023; Xu et al., 2023).

Xu and Durrett (2021) investigated the role of the encoder and decoder components in the BART (Lewis et al., 2020) model in the summarization task. Other studies in summarization use text alignment for corpus creation (Tardy et al., 2020) and detect model hallucinations in summaries using mutual information (van der Poel et al., 2022).

2.2. Eye-gaze studies in NLP Tasks

The Eye-tracking device is a powerful tool to collect eye-gaze data during human cognitive activities. It provides a sequence of screen coordinates of human gaze points with timestamps. Recently, eye-tracking software that uses inexpensive Web cameras has become available(Papoutsaki et al., 2016; Ribeiro et al., 2023). The collected gaze points are clustered into fixations that are collections of close gaze points in terms of both space and time. A fixation consists of the start and end time points and the coordinates of the centroid of the belonging gaze points. The centroid coordinates can be mapped to an object on the screen, e.g., a word in a text.

A bulk of eye-gaze datasets in writing activity has been collected in translation process studies (Carl, 2012a). The collection is well supported by the data collection tool Translog(-II) (Jakobsen, 1999; Schou et al., 2009; Carl, 2012b), which records user's eye-gaze points and keystroke logs. Although Translog was initially designed for the translation study, it can also be used for other writing activities (Sahoo and Carl, 2019). Rodeghero and McMillan (2015) analyzed eye movement patterns for program comprehension through writing in-line code summaries (Rodeghero and McMillan, 2015).

In this research, we use the Translog-II (Carl, 2012b) that records the user's eye movement from the eye-tracker device and keystrokes logs during a writing activity. Translog-II also transforms collected eye-gaze points into a sequence of fixations on words in the text on the screen.

3. Eye-gaze Data for Summarization

We use two eye-gaze datasets collected during human summarization in this study. Table 1 shows the statistics of the datasets.

Dataset	CS19	IELTS33
#Participants	13	11
#Source texts	6	3
#Summaries	26	33
Ave. source length [word]	141	867
Ave. source length [sentence]	6.3	15
Ave. reading time [min]	0.5	4
Ave. writing time [min]	6	14
Reduction rate [%]	80	22

Table 1: Statis of eye-gaze datasets

CS19 Sahoo and Carl (2019) collected eye-gaze data for three writing tasks: text copying, paraphrasing, and summarization in English. We use their summarization data in this study. Thirteen people wrote a total of 26 summaries from six source texts. Four out of six source texts come from news articles. The other two source texts are sociological texts from an encyclopedia. Each summary was written by at least four people.

IELTS33 We also collected eye-gaze data during summarization using Translog-II[1]. We randomly selected three texts from the English proficiency test IELTS that would fit on our screen while excluding texts that were too similar to each other. The texts describe scientific discussions on the effects of noise, 20th-century architecture, and endangered languages, which are 834, 955, and 813 words in length, respectively. We recounted 11 participants, 10 males and one female for the data collection experiment. They are mostly native speakers or near-native speakers of English proficiency. One was a master's-level computer science professional, four were undergraduate students, and seven were PhD students.

The participants used a workstation equipped with an eye tracker (Tobbi Pro X3-120) and Translog-II software running on Windows. The participants use a chinrest, which helps to fix the distance between the eyes and the screen and boosts the eye-tracker's overall accuracy without being

[1]The resource has not been published yet. We consider its publication upon the paper's acceptance

overly intrusive. Three observed summarization sessions were conducted following a brief training phase. Each session began with a calibration phase on Tobii, with brief breaks between sessions. An entire session with three texts takes, on average, 1.5 hours. We collected 33 summaries, 11 for each source text in total.

After the data collection, we noticed four participants had a high rate of vertical errors in the eye tracker. To remedy the vertical errors, we apply an error correction algorithm (Mishra et al., 2012), which vertically shifts the fixations to the nearest line of the previous gaze and discards the vertical jump based on the past and future gaze location. The algorithm was applied to all data, setting the algorithm threshold so that the correction did not affect the other less errorous data. To check the validity of the correction, we manually compare the data before and after the correction.

4. Experimental Setting

In the experiments, we use BART (Lewis et al., 2020) as our primary target architecture as it is widely used in summarization task [2]. We consider three BART variations: the pre-trained BART model (BART-PT[3]), the fine-tuned BART model by the xsum dataset (Narayan et al., 2018) (BART-FT)[4], and its distilled BART model (DistilBART[5]).

We consider four interpretation methods: Input Gradient (Grad), Integrated Gradient (IG) (Shrikumar et al., 2017), Occlusion (Occ) (Zeiler and Fergus, 2014) and Attention (Attn) (Bahdanau et al., 2014). In addition, we consider two baselines: Random, which randomly assigns a token distinct integer representing an importance ranking, and Lead, which assigns a token a rank based on its input position.

As an eye-gaze feature, we adopt fixation count (FC), defined as the number of fixations on a specified object during a specified duration.

We run the summarization models using force decoding to generate human summaries.

5. Macroscopic Analysis

This section analyzes the correlation between the model saliency and fixation counts over the summary generation process. We calculate the word saliency scores by an interpretation method at each step of generating a word of the output summary. For each source text, we conduct word-wise aggregation of the saliency scores across the entire generation steps. As a result, we obtain a saliency vector for the source text in which the dimensions correspond to the word token, and their values indicate the word saliency. We consider two aggregation methods, max and mean. The max aggregation takes the maximum saliency score across all word generation steps as a saliency vector element, while the mean aggregation takes the average saliency scores. Likewise, we aggregate the fixation counts of each word token in the source text by summing up them across the summary writing process to obtain a fixation count (FC) vector for the source text. We consider summary writing to start at the first character input of the summary. We do not distinguish fixations from different participants in the summation.

We also consider a sentence-wise saliency vector and an FC vector in which the vector dimension corresponds to a sentence in the source text, and its value denotes the sentence saliency or sentence fixation counts. The vector values are calculated by averaging the token values in the sentence.

5.1. Discussion for RQ1

To answer RQ1, we calculate Spearman's rank correlation ρ between the saliency vector and FC vector of each source text. We then take their average Spearman's ρ across all source texts by transforming the value into the Fisher's z score and converting it back to ρ value (Myers and Sirois, 2006).

Table 2 shows the correlation between the model saliency and fixation counts averaged over the source texts. Overall, the max aggregation tends to provide higher correlations than the mean aggregation except for Occ. The mean aggregation normalizes the total saliency score by the output summary length. Therefore, it pushes down the aggregated saliency score of words that are highly salient in a few word generation steps. We can conclude that the mean aggregation is inappropriate for the macroscopic analysis. We focus the result by the max aggregation (colored rows) in the following discussion.

Among the interpretation methods, Attn shows stable high correlations, particularly on token-based correlations, followed by the gradient-based methods (Grad and IG). The Lead method shows a significantly high sentence-based correlation for CS19. This high correlation can be explained by the domain bias, i.e. four out of six source texts in CS19 are news articles. In the news domain, the important information tends to be placed in the earlier part of texts by the writing convention in journalism. This explanation is supported by the Lead model's low token-based correlation in CS19 and the low correlation in IELTS33. The source texts

[2] Based on the Huggingface download metrics search Link on 13th Oct 2023

[3] https://huggingface.co/facebook/bart-large

[4] https://huggingface.co/facebook/bart-large-xsum

[5] https://huggingface.co/sshleifer/distilbart-xsum-6-6

Dataset		CS19		IELTS33	
Model+Interp.	Aggr.	token	sent.	token	sent.
Random		.114	.292	.052	.143
Lead		-.541	.762	-.319	.012
BART-PT Grad	max	.360	.307	.401	-.088
	mean	-.092	-.042	.003	-.066
BART-PT IG	max	.299	-.007	.329	.131
	mean	-.018	-.200	.012	.140
BART-PT Occ	max	.009	-.380	-.046	-.317
	mean	.059	.203	.014	-.202
BART-PT Attn	max	.420	.645	.338	.080
	mean	.477	.455	.289	.297
BART-FT Grad	max	.441	.282	.269	-.010
	mean	.227	.207	.050	-.127
BART-FT IG	max	.433	.644	.222	-.087
	mean	.238	-.002	.064	.044
BART-FT Occ	max	.219	-.360	.076	.236
	mean	-.029	.129	.102	.149
BART-FT Attn	max	.545	.251	.395	.058
	mean	.488	.659	.281	.004
DistilBART Grad	max	.319	.362	.222	-.105
	mean	.089	.172	-.009	-.391
DistilBART IG	max	.399	.204	.257	.108
	mean	.113	.045	.032	-.057
DistilBART Occ	max	.002	.513	-.049	NaN
	mean	.117	-.363	.088	.400
DistilBART Attn	max	.510	.546	.401	.020
	mean	.457	.423	.263	-.129

Table 2: Average rank correlation between model saliency and fixation counts

of IELTS33 are scientific articles, which has a different writing style from news articles.

Comparing the two datasets, the correlations in CS19 tend to be higher than those of IELTS33. The difference in source text length, 141 vs 867 words on average, can explain this tendency. In addition, the reduction rate of IELTS33 is four times higher than that of CS19. In CS19, the reduction rate is 80%, which means summarization removes about one sentence from the source text. We expect that the source text and summaries will be very similar, leading to high correlations in both token and sentence-based metrics. On the contrary, the IELTS33 texts are lengthy, and the reduction rate is high, i.e., summaries are one-fourth of the source texts in length. We expect more various operations, such as paraphrasing, splitting sentences, and deleting sentences, applied in the IELTS33 summaries, which makes the alignment between the source text and summaries more difficult. The low correlations of IELTS33, particularly in the sentence-based metric, support this explanation. We also notice the difference in the token-based correlation between the datasets is larger in the fine-tuned models (BART-FT and DistilBART) than in the pre-trained model (BART-PT). This difference can also be explained by the domain bias, as we mentioned above. Since the news articles are dominant in the xsum dataset, the former models successfully fine-tuned which words in the source text to focus on when summarising a news article, which is dominant in CS19 as well.

Regarding the models, we observe consistently high token-based correlations by the BART-FT model in CS19, regardless of the interpretation method. However, we can not observe clear differences among models in the other columns in Table 2.

The answer to RQ1 is that we observe weak correlations between word saliency from the interpretation method and fixation counts on words under some conditions. More concretely, the attention-based interpretation method (Attn) is promising.

5.2. Discussion for RQ2

	Rogue-1	Rogue-2	Rouge-L
CS19			
BART-PT	.450	.181	.273
BART-FT	.299	.079	.192
DistilBART	.303	.073	.191
IELTS33			
BART-PT	.395	.105	.191
BART-FT	.287	.057	.163
DistilBART	.267	.044	.153

Table 3: Average Rouge values of the models

To answer RQ2, we calculate the Rouge scores (Lin, 2004) of the summaries generated by the three summarization models. As there are multiple summaries for a single source text, a Rouge score of the model output is calculated using each human summary as a reference, and the average of these values is used as the evaluation score of the model. Table 3 shows the Rouge scores of the three models. We can see that BART-PT shows the best performance for both datasets. As far as this macroscopic analysis, considering that BART-FT showed notable superiority in CS19, we can not observe the relationship that a well-correlated model to human eye-gaze performs better in summarization. These results lead us to conduct a microscopic analysis in the next section.

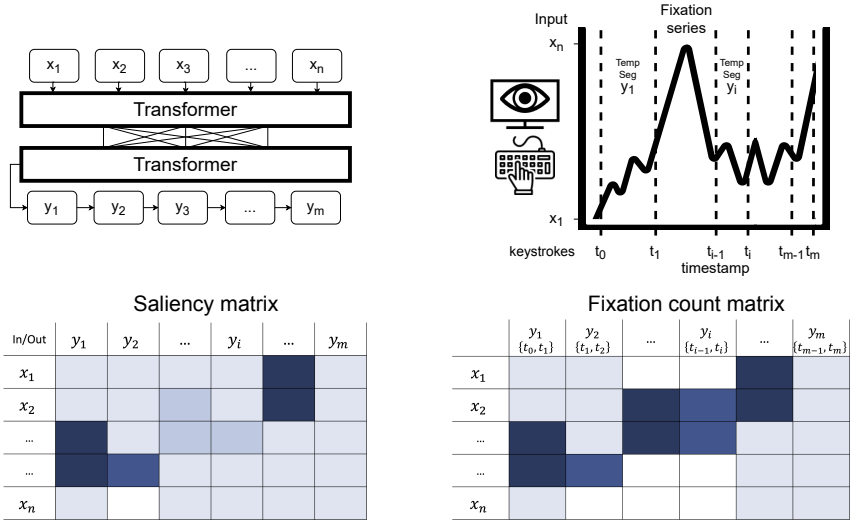

Figure 1: Model saliency matrix (lowe-left) and Fixation count matrix (lower-right). Dense color represents high saliency and frequent fixation counts, respectively.

6. Microscopic Analysis

This section analyzes the relationship between the model saliency and fixation counts on words at each step of generating a word in summaries. Also, we look at their relation and the quality of summaries.

To investigate the relationship between model saliency and fixation counts, we first need to align these two values. Given a source text consisting of n word tokens $\{x_1, x_2, \ldots, x_n\}$, we have a word saliency vector from the interpretation method at each step of generating a summary word y_i, i.e., m vectors in total that constitute a saliency matrix as shown in the lower-left of Figure 1. Likewise, we would like to create a fixation count matrix. However, we must define a duration corresponding to each word in a summary. To define a temporal segment for a word, we consider two methods: fixed-number segmentation (Fix) and keystroke-based segmentation (Key).

The fixed-number segmentation follows the piecewise approximation aggregation (PAA) (Keogh et al., 2001), which segments a time series data by dividing them into temporally equal-sized segments; the value of each segment is calculated by averaging values in the segment. In our case, we divide a sequence of fixation counts for the entire summarization process into m segments of equal duration and sum up the fixation counts in each segment. In this method, however, each segment is not guaranteed to align with the generated word.

To realize a more precise alignment between a temporal segment and a generated word, we introduce keystroke-based segmentation. We define a temporal segment for a generated word as a duration between the first and last key input time points of the word. Participants might edit the word until they finalize it; we include editing time to the temporal segment.

The fixation counts are summed up in each segment. The right of Figure 1 illustrates the keystroke-based segmentation and the resultant fixation count matrix.

Dataset\Seg.	Fix		Key	
Model+Interp.	token	sent.	token	sent.
CS19				
Random	-.0017	.0435	-.0015	.0636
BART-PT Grad	.0381	.1549	.0672	.4267
BART-PT IG	.0160	.0416	.0568	.2957
BART-PT Occ	-.0021	.0113	.0366	-.0007
BART-PT Attn	.1137	.2010	.1796	.4672
BART-FT Grad	.0749	.2821	.1109	.5254
BART-FT IG	.0539	.2800	.0737	.5056
BART-FT Occ	.0160	.0601	.0299	.0941
BART-FT Attn	.1146	.2970	.1836	.5529
IELTS33				
Random	-.0038	-.0068	-.0002	.0184
BART-PT Grad	-.0086	.0780	.0058	.2056
BART-PT IG	-.0020	.0576	.0091	.1897
BART-PT Occ	-.0032	-.0063	.0100	.0183
BART-PT Attn	.0324	-.0495	.0515	.0536
BART-FT Grad	-.0082	.0744	.0059	.1757
BART-FT IG	-.0053	.0448	.0084	.1460
BART-FT Occ	-.0053	-.0313	.0142	.0288
BART-FT Attn	.0242	-.0117	.0407	.0634

Table 4: Average rank correlation at each word generation

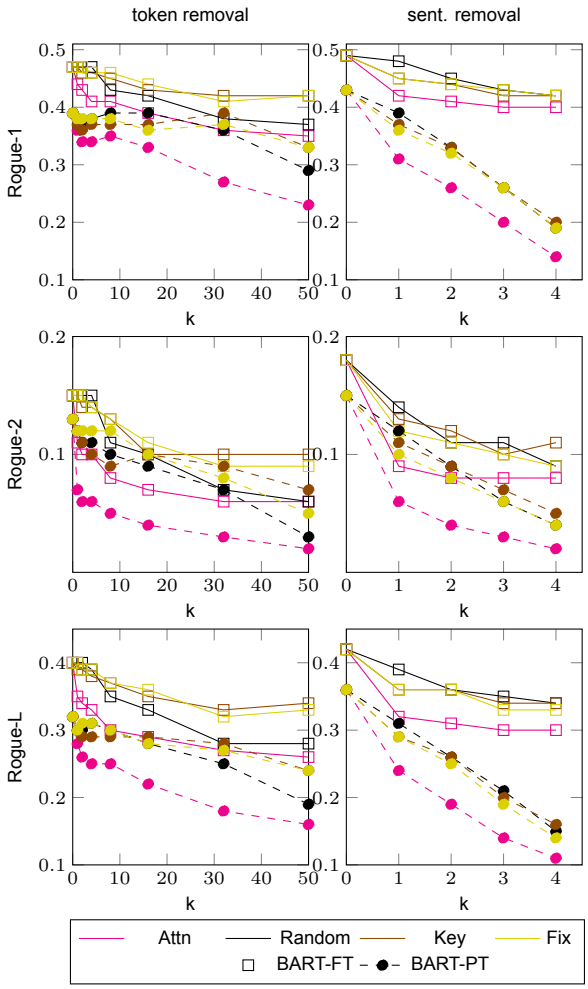

Figure 2: Ablation analysis for CS19

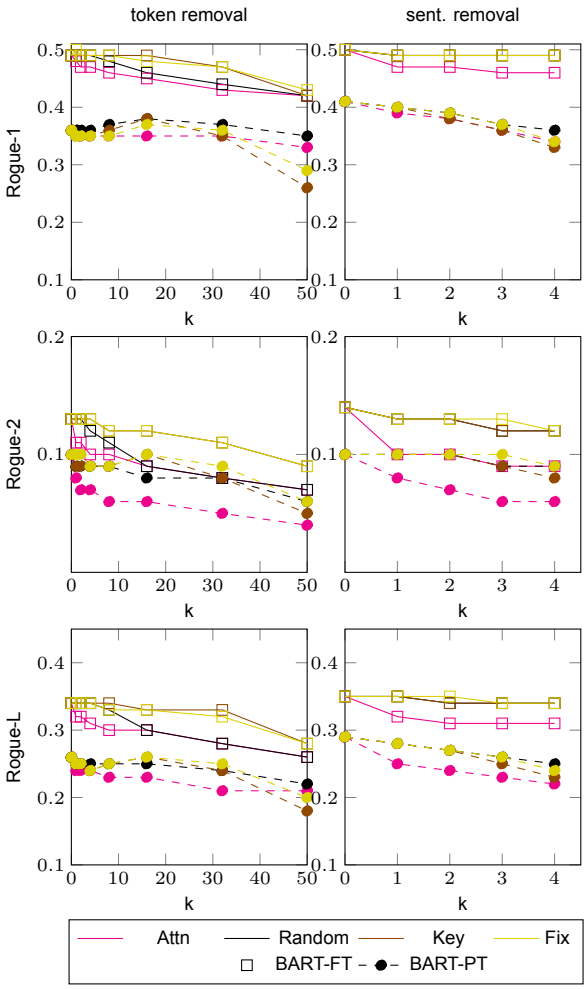

Figure 3: Ablation analysis for IELTS33

6.1. Discussion for RQ1

We calculate the rank correlation between model saliency and human fixation counts at each step of generating a word. Given a source text, we obtain a saliency vector (a column vector in the left bottom matrix in Figure 1) at each step of generating each human summary by force decoding. For a human summary of m word tokens, we have m saliency vectors to construct a saliency matrix shown in the left bottom of Figure 1. Unlike the macroscopic analysis, we can not aggregate the saliency matrices across the participants because the summary length (m) is different depending on each participant. Therefore, we calculate Spearman's rank correlation between each pair of individual saliency vectors (a column in the left bottom of Figure 1) and the corresponding fixation vector (a column in the right bottom). They are averaged over m words in a summary, and further averaged over the participants and source texts. Similar to the macroscopic analysis, we transform the value into the Fisher's z score and convert it back to ρ value (Myers and Sirois, 2006) to calculate the averaged rank correlation.

Table 4[6] shows the average rank correlation between model saliency and fixation counts. Similar to the macroscopic analysis, Attn shows a stable and slightly higher correlation compared to other interpretation methods at the sentence level. However, compared with the result of the macroscopic analysis (Table 2), the correlation coefficients are significantly low, particularly at the token level.

We suspect a single word is too small to capture a relationship between the model saliency and fixation counts. Some larger linguistic units, such as phrases, clauses, and sentences, should have been considered for aggregation.

In most cases, model correlations have a higher correlation to the keystroke-based segmentation (Key) compared to the fixed-number segmentation (Fix). These results show that the model generation process aligns better with keystrokes than the fixed-size duration, which assumes humans write linearly in time without backtracking.

[6]Due to the space limitation, we show only a part of the results. The results of other conditions are similar.

The answer to RQ1 is that we observe weak correlations between word saliency from the interpretation method and fixation counts on the source text sentences under some conditions. Similar to macroscopic analysis, the attention-based interpretation method (Attn) is promising.

6.2. Discussion for RQ2

We apply the input ablation method, which was used to evaluate model faithfulness in NLP tasks (Jacovi and Goldberg, 2020). The method is also applied for evaluating saliency in NLP models (DeYoung et al., 2020; Xu and Durrett, 2021). Xu and Durrett (2021) used the model loss values against adding or removing salient parts of the input, such as token(s) or sentence(s), to investigate the relationship between model saliency and model performance. We use the Rouge score of the generated summaries instead of the model loss because Rouge scores indicate the summary quality more directly. We calculate the Rouge scores the same as in 5.2, i.e. using human summaries as the reference summary.

Following Xu and Durrett (2021), we replace the top k salient words with a special mask token at each step of generating a word, and calculate the Rouge score of the resultant summary. We also conduct the word ablation based on the fixation count and compare the change of Rouge scores between model saliency and human fixation counts. In addition, we conduct sentence ablation, where top k salient sentences are simply removed from the input.

Figure 2 and 3 show the Rouge score (y-axis) against the number of removed words/sentences at each generation step (x-axis). To avoid the diagrams becoming complicated, we show the results of the model saliency-based ablation (Attn), a baseline (Random), and the fixation-based ablation (Key and Fix). Between the datasets, we notice the ablation impacts the Rouge scores more mildly in IETLS33 than CS19. Particularly, the difference is significant in the sentence ablation. This happens due to longer source texts in the IELTS33 dataset.

BART-FT shows consistently higher Rouge scores than BART-PT, which is the opposite result of the summary generation without force decoding (Table 3). However, this result conforms with the higher correlation of BART-FT than BART-PT in the macroscopic analysis.

We do not observe much difference between the temporal segmentation methods: fixed-number segmentation (Fix) and keystroke-based segmentation (Key) in these graphs. Surprisingly, Random often shows lower Rouge scores than the fixation-based ablation.

The saliency-based ablation (Attn) shows much lower Rouge scores than Random. This observation suggests the summarization models look at different parts of the input than humans do when generating a summary. To summarize, our tentative answer to RQ2 in this microscopic analysis is negative.

7. Conclusion and Future Work

In this paper, we proposed a novel framework for analyzing summarization models by comparing them to the human summarization process with eye movement as a proxy. Our framework comprises macroscopic and microscopic analysis between model saliency and human gaze data. In macroscopic analysis, we compared the model saliency and eye-gaze at the input level. In microscopic analysis, we compared the model saliency and eye-gaze at each output token. To align model saliency and eye-gaze information at every token generation, we introduced the keystroke-based segmentation for the time series of fixations.

We answered two key research questions using the framework. RQ1: *Does the input word saliency from interpretation methods in summarization models conform with human eye-gaze features during summarization?* RQ2: *How do the model saliency and human visual attention from eye movement affect model performance?*.

According to the correlations between model saliency scores and human fixation counts in the macroscopic and microscopic analyses, our answer to RQ1 is partially yes, particularly for the attention-based saliency scores. Our ablation analysis showed that removing important words according to the human gaze did not degrade performance as much as the removal by the model saliency. Thus, the answer to RQ2 is that the summarization models look at different input words than humans in generating summaries.

Our experiments compared a limited number of interpretation methods; other methods should be also considered. For instance, a recent study by Eberle et al. (2022) reported the attention flow method shows higher correlations to human eye movements in sentiment analysis and relation extraction tasks, although they do not have a writing phase. In addition, we could consider more varied configurations for text generation, in terms of different architectures, e.g. encoder-decoder models vs. decoder-based models, different generation approaches, e.g. autoregressive vs. non-autoregressive, and different decoding strategies, e.g. greedy vs beam-search.

Last but not least, collecting more eye-gaze data in the summarization task is indispensable. However, it is more time-consuming and expensive than that for reading tasks. Participants have to spend more time and effort to read and produce

a decent summary. Using synthesized eye-gaze data (Sood et al., 2020b) will be a reasonable option for scaling up the eye-gaze data and analyzing the writing behavior of humans and machines.

Our approach can be applied to other text generation tasks such as paraphrasing and machine translation. The main difference between these tasks and text summarization is that the source and target texts are aligned in their length. There have already been studies on eye movement analysis during human translation (Carl et al., 2008; Jakobsen, 2011; Carl and Kay, 2012). Our method can help in comparing the attention of humans and machines in translation.

8. Ethical Statement

The data collection experiments for IELTS33 were reviewed and approved by the Ethical Review Committee of the author's university in advance. Prior to the experiment, the purpose and method of data collection, and usage of the collected data were explained to the participants, and their consent to participate in the experiment was obtained.

9. Bibliographical References

David Alvarez-Melis and Tommi Jaakkola. 2017. A causal framework for explaining the predictions of black-box sequence-to-sequence models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 412–421, Copenhagen, Denmark. Association for Computational Linguistics.

Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. 2018. Towards better understanding of gradient-based attribution methods for deep neural networks.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Yonatan Belinkov, Sebastian Gehrmann, and Ellie Pavlick. 2020. Interpretability and analysis in neural NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 1–5, Online. Association for Computational Linguistics.

Michael Carl. 2012a. The CRITT TPR-DB 1.0: A database for empirical human translation process research. In *Workshop on Post-Editing Technology and Practice*, San Diego, California, USA. Association for Machine Translation in the Americas.

Michael Carl. 2012b. Translog-II: a program for recording user activity data for empirical reading and writing research. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 4108–4112, Istanbul, Turkey. European Language Resources Association (ELRA).

Michael Carl, Arnt Lykke Jakobsen, and Kristian T.H. Jensen. 2008. Modelling human translator behaviour with user-activity data. In *Proceedings of the 12th Annual Conference of the European Association for Machine Translation*, pages 21–26, Hamburg, Germany. European Association for Machine Translation.

Michael Carl and Martin Kay. 2012. Gazing and typing activities during translation: A comparative study of translation units of professional and student translators. *Meta: Translators' Journal*, 56:952–975.

Charles Clifton, Adrian Staub, and Keith Rayner. 2007. Chapter 15 - eye movements in reading words and sentences. In Roger P.G. Van Gompel, Martin H. Fischer, Wayne S. Murray, and Robin L. Hill, editors, *Eye Movements*, pages 341–371. Elsevier, Oxford.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Shuoyang Ding, Hainan Xu, and Philipp Koehn. 2019. Saliency-driven word alignment interpretation for neural machine translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 1–12, Florence, Italy. Association for Computational Linguistics.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning.

Oliver Eberle, Stephanie Brandl, Jonas Pilot, and

Anders Søgaard. 2022. Do transformer models show similar attention patterns to task-specific human gaze? In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4295–4309, Dublin, Ireland. Association for Computational Linguistics.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of neural models make interpretations difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3719–3728, Brussels, Belgium. Association for Computational Linguistics.

Chaoyu Guan, Xiting Wang, Quanshi Zhang, Runjin Chen, Di He, and Xing Xie. 2019. Towards a deep and unified understanding of deep neural models in NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2454–2463. PMLR.

Shilin He, Zhaopeng Tu, Xing Wang, Longyue Wang, Michael Lyu, and Shuming Shi. 2019. Towards understanding neural machine translation with word importance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 953–962, Hong Kong, China. Association for Computational Linguistics.

Nora Hollenstein and Lisa Beinborn. 2021. Relative importance in sentence processing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 141–150, Online. Association for Computational Linguistics.

Nora Hollenstein, Antonio de la Torre, Nicolas Langer, and Ce Zhang. 2019. CogniVal: A framework for cognitive word embedding evaluation. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 538–549, Hong Kong, China. Association for Computational Linguistics.

Fariz Ikhwantri, Jan Wira Gotama Putra, Hiroaki Yamada, and Takenobu Tokunaga. 2023. Looking deep in the eyes: Investigating interpretation methods for neural models on reading tasks using human eye-movement behaviour. *Information Processing & Management*, 60(2):103195.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205, Online. Association for Computational Linguistics.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. Association for Computational Linguistics.

Arnt Lykke Jakobsen. 1999. Logging target text production with translog. *Copenhagen studies in language*, pages 9–20.

Arnt Lykke Jakobsen. 2011. Tracking translators' keystrokes and eye movements with translog.

Marcel Adam Just and Patricia A Carpenter. 1980. A theory of reading: from eye fixations to comprehension. *Psychological review*, 87 4:329–54.

Eamonn J. Keogh, Selina Chu, David Hart, and Michael J. Pazzani. 2001. An online algorithm for segmenting time series. In *Proceedings of the 2001 IEEE International Conference on Data Mining*, ICDM '01, page 289–296, USA. IEEE Computer Society.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2016. Visualizing and understanding neural models in NLP. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 681–691, San Diego, California. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Zachary C. Lipton. 2018. The mythos of model interpretability: In machine learning, the concept

of interpretability is both important and slippery. *Queue*, 16(3):31–57.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3730–3740, Hong Kong, China. Association for Computational Linguistics.

Yu Lu, Jiajun Zhang, Jiali Zeng, Shuangzhi Wu, and Chengqing Zong. 2022. Attention analysis and calibration for transformer in natural language generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:1927–1938.

Abhijit Mishra, Michael Carl, and Pushpak Bhattacharyya. 2012. A heuristic-based approach for systematic error correction of gaze data for reading. In *Proceedings of the First Workshop on Eye-tracking and Natural Language Processing*, pages 71–80, Mumbai, India. The COLING 2012 Organizing Committee.

Leann Myers and Maria J. Sirois. 2006. *Spearman Correlation Coefficients, Differences between*. John Wiley & Sons, Ltd.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Alexandra Papoutsaki, Patsorn Sangkloy, James Laskey, Nediyana Daskalova, Jeff Huang, and James Hays. 2016. Webgazer: Scalable webcam eye tracking using user interactions. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3839–3845. AAAI.

Keith Rayner. 1998. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124 3:372–422.

Keith Rayner. 2009. Eye movements and attention in reading, scene perception, and visual search. *The Quarterly Journal of Experimental Psychology*, 62(8):1457–1506.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Tiago Ribeiro, Stephanie Brandl, Anders Søgaard, and Nora Hollenstein. 2023. Webqamgaze: A multilingual webcam eye-tracking-while-reading dataset.

Daniel C. Richardson, Rick Dale, and Michael J. Spivey. 2007. Eye movements in language and cognition: A brief introduction. In Monica Gonzalez-Marquez, Irene Mittelberg, Seana Coulson, and Michael J. Spivey, editors, *Methods in Cognitive Linguistics*, pages 323–344. John Benjamins.

Paige Rodeghero and Collin McMillan. 2015. An empirical study on the patterns of eye movement during summarization tasks. In *2015 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, pages 1–10.

Debasish Sahoo and Michael Carl. 2019. Lexical representation & retrieval on monolingual interpretative text production. In *Proceedings of the Second MEMENTO workshop on Modelling Parameters of Cognitive Effort in Translation Production*, pages 14–16, Dublin, Ireland. European Association for Machine Translation.

Lasse Schou, Barbara Dragsted, and Michael Carl. 2009. Ten years of translog. *Copenhagen studies in language*, pages 37–48.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3145–3153. JMLR.org.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps.

Ekta Sood, Simon Tannert, Diego Frassinelli, Andreas Bulling, and Ngoc Thang Vu. 2020a. Interpreting attention models with human visual attention in machine reading comprehension. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 12–25, Online. Association for Computational Linguistics.

Ekta Sood, Simon Tannert, Philipp Müller, and Andreas Bulling. 2020b. Improving natural language processing tasks with human gaze-guided neural attention. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1–15.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2020. Learning to summarize from human feedback. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328. JMLR.org.

Joël Tang, Marina Fomicheva, and Lucia Specia. 2023. Reducing hallucinations in neural machine translation with feature attribution.

Paul Tardy, David Janiszek, Yannick Estève, and Vincent Nguyen. 2020. Align then summarize: Automatic alignment methods for summarization corpus creation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6718–6724, Marseille, France. European Language Resources Association.

Keyon Vafa, Yuntian Deng, David Blei, and Alexander Rush. 2021. Rationales for sequential predictions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10314–10332, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Laurens van der Maaten and Geoffrey E. Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605.

Liam van der Poel, Ryan Cotterell, and Clara Meister. 2022. Mutual information alleviates hallucinations in abstractive summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5956–5965, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shikhar Vashishth, Shyam Upadhyay, Gaurav Singh Tomar, and Manaal Faruqui. 2019. Attention interpretability across nlp tasks.

Jesse Vig. 2019. A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42, Florence, Italy. Association for Computational Linguistics.

Elena Voita, Rico Sennrich, and Ivan Titov. 2021. Analyzing the source and target contributions to predictions in neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1126–1140, Online. Association for Computational Linguistics.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

Jiacheng Xu and Greg Durrett. 2021. Dissecting generation modes for abstractive summarization models via ablation and attribution. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6925–6940, Online. Association for Computational Linguistics.

Weijia Xu, Sweta Agrawal, Eleftheria Briakou, Marianna J. Martindale, and Marine Carpuat. 2023. Understanding and detecting hallucinations in neural machine translation via model introspection. *Transactions of the Association for Computational Linguistics*, 11:546–564.

Matthew D. Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *Computer Vision – ECCV 2014*, pages 818–833, Cham. Springer International Publishing.

10. Language Resource References

Debasish Sahoo and Michael Carl. 2019. *Lexical Representation & Retrieval on Monolingual Interpretative text production*. The Center for Research in Translation and Translation Technology (CRITT). European Association for Machine Translation. [link].